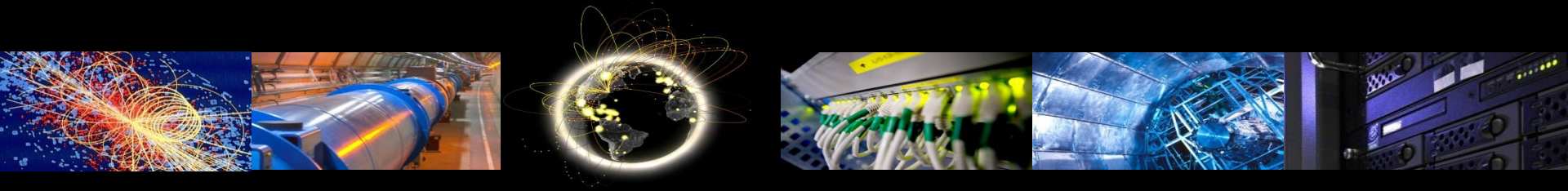
vCHEP 2021 Facilities and Networks Summary

Grid Deployment Board

July 14th 2021



Session 1

- Ethernet evaluation in data distribution traffic for the LHCb filtering farm at CERN
- Systematic benchmarking of HTTPS third party copy on 100Gbps links using XRootD
- NOTED: a framework to optimise network traffic via the analysis of data from File Transfer Services
- Benchmarking NetBASILISK: a Network Security Project for Science
- Proximeter: CERN's detecting device for personnel

Session 1

Ethernet evaluation in data distribution traffic for the LHCb filtering farm at CERN; *Rafał Dominik Krawczyk* (CERN)

- Challenge: LHCb hardware trigger doesn't profit from higher luminosity - go trigger-less instead
- Commisioning (allegedly) largest real-time acquisition system in world (32 Tb/s)
- Can second stage of Event Builder perform well-enough with Ethernet RDMA over Converged Ethernet (RoCE) v2?

- Sufficient real-time operation of the LHCb-like traffic over Ethernet
- Hybrid InfiniBand + Ethernet EB applicable
- Proof of concept made → LHCb EB can handle 32 Tbit/s readout
- Ethernet RoCE v2 evolution followed → full EB for Run 4 ?

Session 1

Systematic benchmarking of HTTPS third party copy on 100Gbps links using XRootD; *Aashay Arora* (UCSD)

- To systematically execute TPC transfers both with GridFTP and XRootD-HTTPS over 100Gbps links and compare their performance
- Study the sensitivity of these protocols to latency (RTT) between servers
- Methodology using containers deployed on the Pacific Research Platform, a worldwide distributed Kubernetes cluster

- HTTPS consistently overperforms GridFTP (about 30% better transfer rates on average) given same setup
- Able to achieve upto 45 Gbps using XRootD-HTTPS but we are still significantly below the throughput of the same hardware using ethr
- Throughput decreases significantly as latency increases.

Session 1

NOTED: a framework to optimise network traffic via the analysis of data from File Transfer Services; *Edoardo Martelli* (CERN)

- NOTED aims to detect the start of large data transfers and to predict their duration
- Ultimate intention is to reduce the transfer duration, improve the effective transfer bandwidth for users, improve the efficient utilisation of available networks
- Transfer Broker -> Network Intelligence -> Software Defined Network Controller
- NOTED has been tested to identify real transfers and apply real network optimizations

- Analysing data from transfer services can allow the understanding and the prediction of network traffic
- NOTED can use these predictions to reconfigure the network and deliver faster data transfers and more efficient network utilization

Session 1

Benchmarking NetBASILISK: a Network Security Project for Science; *Jem Guhit* (University of Michigan)

- Enable movement of science data to and from servers while receiving protection against threats; firewalls block all data, which impedes scientific research
- Services in Science DMZs are more vulnerable; expand Science DMZs by enabling Internet measurement and monitoring without distorting the resulting data
- NetBASILISK observes and analyzes traffic without interfering with the network
- Build a framework composed of benchmark and environment monitoring application; assist in determining if poor transfer results are due to NetBASILISK

- Framework Script is complete, automated, and output sent to the Humio database
- An Independent Environment Service is being developed to observe variability for a time-range to define the baselines for a busy, medium, and free network
- Next phase would focus on analysis of benchmark and environment metadata

Session 1

Proximeter: CERN's detecting device for personnel; Christoph Merscher (CERN)

- Protect employees to counteract the spread of the COVID-19 virus as much as possible
- Need to detect an encounter of less than 2m
- Hardware proximeter sensor developed at CERN
- LoRaWAN is a Low Power WAN protocol; network in place at CERN

- LoRaWAN network and the proximeter have been designed together to cope with the project requirements
- Makes it possible to trace who a person encountered anonymously
- Only medical service can see all information
- Have given visibility to the LoRaWAN project

Session 2

- Deploying a new realtime XRootD-v5 based monitoring framework for GridPP
- Towards Real-World Applications of ServiceX, an Analysis Data Transformation System
- Anomaly detection in the CERN cloud infrastructure
- Reaching new peaks for the future of the CMS HTCondor Global Pool
- Research and Evaluation of RoCE in IHEP Data Center

Session 2

Deploying a new realtime XRootD-v5 based monitoring framework for GridPP; Rob Currie (Universty of Edinburgh)

  - Availability/Reliability; XCache storage; XCache networking performance stats; cache effectiveness: *Does site/job-efficiency/reliability improve*?
  - New system based on OSG XRootD collector
  - Monitoring services at two different sites in GridPP

- Built upon the OSG collector to collect additional caching metrics from distributed XCache instances.
- Redesigned Edinburgh ELK+ monitoring stack to be more resilient/scalable/secure.
- Taken preliminary measurements of metrics such as cache-efficiency/effectiveness using new monitoring stack.
- Demonstrated that deploying an XCache at a site can offer improved performance.
- Examine potential performance/scalability improvements in both XCache and monitoring system

Session 2

Towards Real-World Applications of ServiceX, an Analysis Data Transformation System; *KyungEon Choi* (University of Seoul)

- A scalable HEP event data extraction, transformation and delivery system delivering columnar data on demand
- A building block to construct an analysis workflow
- Showed an example of ServiceX into an existing analysis workflow: TRExFitter
- Optimise delivery of data and reduce workflow durations

- Invitation to try ServiceX: https://servicex.readthedocs.io/en/latest/

Session 2

Anomaly detection in the CERN cloud infrastructure; *Stiven Metaj* (CERN)

- CERN Cloud Infrastructure: ~8k bare-metal servers serving ~35k VMs
- Goals: automate the Anomaly Detection task, discover misbehaviors earlier, consider metrics correlation
- Two approaches: change detection in a single host; outlier in cloud hostgroup
- Implementation and Algorithms, Evaluation and Results

- Designed an Anomaly Detection System with Expert Feedback
- First quantitative evaluation of unsupervised algorithms looks promising for their adoption in the Anomaly Detection for the CERN Cloud use case
- Defined a procedure to annotate and collect time-series datasets

Session 2

Reaching new peaks for the future of the CMS HTCondor Global Pool; ***Antonio Perez-Calero Yzquierdo*** **(CIEMAT)**

- CMS pool of resources under continuous growth: 3x in 5 years
- Implications from HL-LHC challenge: accelerated CPU requirements
- Scalability tests 2021

- The Submission Infrastructure is a stable and performant piece of CMS Computing, continuously being reviewed, upgraded and expanded, in particular prevent scalability bottlenecks
- The Q1 2021 scalability tests of the CMS Global Pool:
- Continued tuning of the infrastructure settings
- Observed improved performance of the collector running on a Physical Machine instead of a VM
- Bottlenecks identified in the collector capacity to process slot updates and memory use in the Central Manager host
- Exceeded the significant value of 500k concurrently running jobs

Session 2

Research and Evaluation of RoCE in IHEP Data Center; *Shan Zeng* (IHEP)

- HPC requires high performance network with many features; High bandwidth; Low latency; Zero package loss; Stable; Scalable; Flexible; Manageable
- Compare InfiniBand to RoCE; RoCEv2 favourable in terms of Ecosystem, Configuration and Cost

- IHEP started to research and evaluate RoCE in the end of last year
- Basic MPI benchmark tests
- RoCE performs slightly better than IB network in both point-to-point and collective tests except for a static latency test

Session 3

- Updates on usage of the Czech national HPC center
- Exploitation of the MareNostrum 4 HPC using ARC-CE
- Exploitation of HPC Resources for data intensive sciences
- Finalizing Construction of a New Data Center at BNL
- Designing the RAL Tier-1 Network for HL-LHC and Future data lake

Session 3

Updates on usage of the Czech national HPC center; *Michal Svatos* (Czech Academy of Sciences)

- ATLAS distributed computing opportunistically using resources of the Czech national HPC center IT4Innovations through Czech Tier2 praguelcg2
- Job submission system (push model); suitable for restrictive environment where only ssh connection through login nodes is available ,i.e. not suitable for data intensive workloads
- Migration to ARC-CE version 6 fixed problems with synchronization of accounting records to APEL
- Connection via sshfs was a bottleneck, improved in v3.7 with `max_conns`; with 10 connections, no link saturation
- Containerization; fat containers; all the necessary software and condition data of one release (~300k files, ~30 GB)
- Long jobs; new setup to run multiple payloads in one pilot job is tested and prepared for newest HPC
- HPCs provided almost half of the wallclock of praguelcg2 in 2020

Session 3

Exploitation of the MareNostrum 4 HPC using ARC-CE; *Andrés Pacheco Pages* (IFAE)

- Primary motivation for integrating the HPC centers in Spain LHC computing is to reduce the cost and take advantage of the new massive computing infrastructures.
- By far, the most powerful machine is MareNostrum 4, located in the Barcelona Supercomputing Center (BSC).
- Flexible structure where the HPC center is one more resource that is added behind the existing Tier1 and Tier2 centers
- Main challenge is the lack of external connectivity at MareNostrum4 computing nodes; focus on the results obtained with ATLAS and the ARC-CE.
- Workflow at MareNostrum4; validated Singularity images with all the software and data preloaded

- ATLAS Simulation jobs integrated into the MareNostrum 4
- BSC has included the LHC computing in the list of strategic projects.
- Expect that transition to MareNostrum 5 can be straightforward with 17 times more computing power in 2021
- Still need grid computing for the LHC; many workflows can't run at BSC for lack of connectivity; need to store, distribute and archive to tape the data

Session 3

Exploitation of HPC Resources for data intensive sciences; *David Southwick* (CERN)

- Exascale HPC machines will provide processing capacities similar to or greater than the entire compute grid; common challenges drive HPC adoption
- Porting experiment workloads to GPUs
- HEP Benchmark Suite extended for HPC
- Data Access: Exascale challenge; Efficient usage of storage systems on site

- To meet a looming computational resource gap, CERN must evolve its computing platform to leverage heterogeneous computing and HPC systems
- The DEEP-EST project proved to be an invaluable platform for collaboration with HPC experts from other sciences/centers
- Developing benchmarking on HPC Next Steps; continue to work with run3 heterogeneous workloads as they become available
- Data Access / HPC:In the coming months, we will leverage on the HPC collaboration with PRACE to demonstrate scale, as well as focus on data ingress/egress

Session 3

Finalizing Construction of a New Data Center at BNL; *Alexandr Zaytsev* (BNL)

- New data center is designed & being constructed for the SDCC Facility in B725 in FY19-21 period; migration of most of the DISK and all the CPU resources to the new data center in FY21-23
- Single large data hall for CPU & DISK resources (Main Data Hall) divided into 2 aisles
- 188 rack positions to become available starting from 2021Q3
- Network Room: Capable of hosting all passive and active network equipment to serve the fully built out B725 datacenter
- Tape Room: Capable of hosting up to 6x 20k slot libraries; in FY21-30 could mean placing up to 11 EB of uncompressed data on TAPE in this area alone by FY30 given the expectations of the LTO technology evolution
- The occupancy of the B725 datacenter for ATLAS is expected to start in June 2021, and occupancy for all tenants – in July 2021
- Early occupancy shifted from 2021Q1 to 2021Q3 as a result of COVID-19
- Scaling up of the B725 data center to 4.8MW of total IT payload capacity is expected in FY25 to address the growing needs of HEP/NP experiments involved

Session 3

Designing the RAL Tier-1 Network for HL-LHC and Future data lake; *Alastair Dewhurst* (UKRI STFC)

- The RAL Tier-1 core network was in need of replacement; needed upgrade the LHCOPN connection to at least 100Gb/s; replace old hardware; new network for new Tape Service; join the LHCONE
- Funds for a major upgrade secured in September 2020
- RAL throughput requirements grow from 100Gb/s in 2023; 200 in 2025; 500 in 2027 (minimum, 800 flexible)
- Everything must support IPv6; must join LHCONE; must be future proof (easily scale up bandwidth and take part in future network activities)
- Storage Requirements - Ceph: Aim to provide ~1Gb/s/HDD
- Storage Requirements - Tape: Aim to provide 200Gb/s
- CPU Requirements: ~0.5MB/s per job slot
- Network infrastructure; 1600Gb/s link to Super Spine

- The RAL Tier-1 has nearly completed a major upgrade to its network
- Assuming flat cash, during LS3 an upgrade to 400Gb/s should be possible
- Providing non-blocking internal networking is a minor cost compared to the external networking

Session 4

- Harnessing HPC resources for CMS jobs using a Virtual Private Network
- Exploitation of network-segregated CPU resources in CMS
- WLCG Token Usage and Discovery
- Secure Command Line Solution for Token-based Authentication
- A Unified approach towards Multi-factor Authentication(MFA)

Session 4

Harnessing HPC resources for CMS jobs using a Virtual Private Network; *Ben Tovar* (University of Notre Dame)

- Resources available (e.g. HPC's) for CMS may be behind firewalls
- How to connect these resources to CMS production infrastructure
- How to provide network resource limit guarantees
- Use DMZs to link HPC resources to CMS; DMZ provides isolated interface to outside network
- Requirements; jobs must run without admin privileges; (CMS) job description should remain unchanged; network usage limits guarantees should be given and enforced
- Construct a Virtual Private Network (VPN); VPN server runs in the DMZ and provides IPs, DNS, and routes to VPN clients; All network traffic of the compute node is managed by the VPN
- Possible uses; Implementation of VPN client and server

- Good performance with multiple concurrent transfers/jobs

Session 4

Exploitation of network-segregated CPU resources in CMS; *Antonio Delgado Peris* (CIEMAT)

- BSC - Barcelona Supercomputing Center; Current MareNostrum 4, MareNostrum5 in 2022
- BSC imposes very restrictive network connectivity conditions; major obstacle for CMS workloads
- Discuss pre-placing whole CVMFS tree at BSC (+ updates); Copying required condition data files: Provides greater flexibility
- Synchronous movement of data (coupled to jobs) preferred; for stage-out, let BSC jobs copy output to per-job area on share filesystem; HTCondor starter at PIC copies whole paths to PIC storage; analogous method could be used for stage-in
- Very successful proof of concept with custom singularity images
- Copied whole cms.cern.ch repository, 12.6TB, 183M files (37M de-duplicated); took ~2 weeks to complete

- Demonstrated BSC can run CMS simulation at scale
- CVMFS repo copy to BSC gives flexibility
- Data staging procedure designed and implemented

Session 4

WLCG Token Usage and Discovery; *Tom Dack* (UKRI STFC)

- WLCG Authorization (AuthZ) WG membership includes current major users of tokens in HEP: INDIGO IAM; EGI Check-in; SciTokens; dCache; ALICE
- Pilot project development work supported by AARC; EOSC-hub; EOSC-pilot
- Stick to industry R&E and industry stands wherever possible
- Collaboration with DOMA WG
- Timeline of progress; documentation; WLCG token: JWT distributed over OAuth2.0 Protocol which contains identity and authz information from issuer (VO)
  - Token schema v1.0
- Many tools will rely on tokens being stored in the local environment
  - Token discoverability specification v1.0
- Rucio-FTS-SE flow; token request and exchange between Rucio, FTS and individual SEs

Session 4

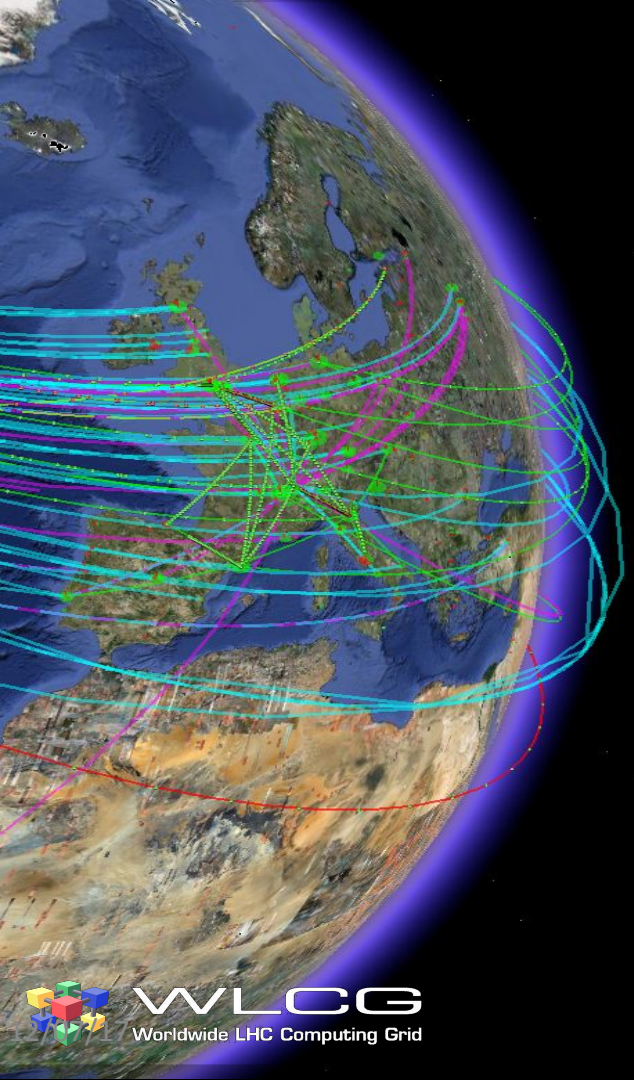
Secure Command Line Solution for Token-based Authentication; *Dave Dykstra* (Fermilab)

- Oauth2/Open ID Connect (OIDC) and JSON Web Tokens (JWTs) - good, but introduce challenges; OIDC assumes web browsers vs command line; need a new way to renew tokens for grid jobs
- WLCG Authz WG considered oidc-agent, but it wasn't a good match (didn't solve MyProxy replacement, not as user friendly as could be)
- Solution: Hashicorp Vault with `htgettoken`
- OIDC flow; htvault-config configuration package
- Capability sets, issuers, and roles; configure Vault to request a scope which the token issuer translates into a set of capability scopes: Groups correspond to VOs and roles within those VOs
- htgettoken and Vault have been integrated into HTCondor
- Token flow with HTCondor and Vault
- Support for "robot" (unattended) operation

Session 4

A Unified approach towards Multi-factor Authentication (MFA); *Masood Zaran* (BNL)

- Adoption of Multi-factor Authentication (MFA) became inevitable at BNL SDCC
- Web applications protected by Keycloak (SSO) MFA, other service components rely on FreeIPA (LDAP) MFA
- Technically satisfies requirements; users need to manage multiple tokens
- Keycloak can't consume FreeIPA based OTP tokens (design limitation); Keycloak & FreeIPA stores user OTP tokens separately, tokens are not interchangeable
- Need to provide users a single OTP which can be used everywhere including other services like mail or future new services
- PrivacyIDEA: Enterprise ready opensource modular authentication system; comes with a variety of MFA tokens and features built-in
- Overview of PrivacyIDEA deployment
- Keycloak enrollment using PrivacyIDEA; Keycloak login using PrivacyIDEA; PrivacyIDEA login / roles
- Customized policies used to restrict authorization; Customization of Web UI access; Restrict specific set of users from logging in all together



Questions?